1

Technical Report 747

The Computerized Adaptive Screening Test (CAST): An Examination of Test Validity and Test Fairness

Deirdre J. Knapp, Rebecca M. Pliske, and
Timothy W. Elig

Selection and Classification Technical Area
Manpower and Personnel Research Laboratory

U. S. Army
Research Institute for the Behavioral and Social Sciences

June 1987

Approved for public release; distribution unlimited.

U. S. ARMY RESEARCH INSTITUTE

FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency under the Jurisdiction of the

Deputy Chief of Staff for Personnel

EDGAR M. JOHNSON Technical Director	WM. DARRYL HENDERSON COL, IN Commanding

Technical review by

Clessen J. Martin
Elizabeth P. Smith

NOTICES

DISTRIBUTION: Primary distribution of this report has been made by ARI. Please address correspondence concerning distribution of reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, ATTN: PERI-POT, 5001 Eisenhower Ave., Alexandria, Virginia 22333-5600.

FINAL DISPOSITION: This report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

A186 999

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS
BEFORE COMPLETING FORM

1. REPORT NUMBER ARI Technical Report 747	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) THE COMPUTERIZED ADAPTIVE SCREENING TEST (CAST): AN EXAMINATION OF TEST VALIDITY AND TEST FAIRNESS		5. TYPE OF REPORT & PERIOD COVERED Interim Report January 1985-March 1986 6. PERFORMING ORG. REPORT NUMBER --
7. AUTHOR(s) Deirdre J. Knapp, Rebecca M. Pliske, and Timothy W. Elig		8. CONTRACT OR GRANT NUMBER(s) --
9. PERFORMING ORGANIZATION NAME AND ADDRESS U.S. Army Research Institute for the Behavioral and Social Sciences 5001 Eisenhower Avenue, Alexandria, VA 22333-5600		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 2Q263731A792 2.2.1.H.3/2.2.1
11. CONTROLLING OFFICE NAME AND ADDRESS U.S. Army Research Institute for the Behavioral and Social Sciences 5001 Eisenhower Avenue, Alexandria, VA 22333-5600		12. REPORT DATE June 1987 13. NUMBER OF PAGES 32
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) --		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

--

18. SUPPLEMENTARY NOTES

--

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Recruiting
Computerized Adaptive Test (CAT)
Computerized Adaptive Screening Test (CAST)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The Computerized Adaptive Screening Test (CAST) is used by Army recruiters to predict prospective applicants' (i.e., prospects') performance on the Armed Forces Qualification Test (AFQT). CAST performance data were collected from 60 recruiting stations across the country throughout calendar year 1985. These data were matched to applicant tapes from Military Entrance Processing Stations (MEPS) to obtain AFQT scores and relevant demographic information. Data analyses indicate that CAST is quite good at predicting AFQT scores for

(continued)

ARI Technical Report 747

20. (continued)

the entire sample and for examinees grouped by sex and race (black or white). When corrected for restriction in range, the cross-validated validity estimate based on the whole sample is .86. Race and sex differences in prediction exist, but these differences are minor and they correspond to those differences found with most cognitive ability tests. CAST's accuracy at predicting subsequent classification into important AFQT categories (i.e., 1-3A and 1-3B) is also discussed.

Technical Report 747

The Computerized Adaptive Screening Test (CAST): An Examination of Test Validity and Test Fairness

Deirdre J. Knapp, Rebecca M. Pliske, and Timothy W. Elig

Selection and Classification Technical Area
Lawrence M. Hanser, Chief

Manpower and Personnel Research Laboratory
Newell K. Eaton, Director

U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES
5001 Eisenhower Avenue, Alexandria, Virginia 22333-5600

Office, Deputy Chief of Staff for Personnel
Department of the Army

June 1987

Army Project Number
2Q263731A792

Manpower and Personnel

Approved for public release; distribution unlimited.

FOREWORD

The Army faces a continuing demand to meet recruiting quantity and quality goals. Recent advances in computer technology and psychometric theory have made possible a new type of assessment technique, called computerized adaptive testing (CAT), that can provide accurate ability estimates based on relatively few test items. The Computerized Adaptive Screening Test (CAST) was designed to provide an estimate of a prospect's Armed Forces Qualification Test (AFQT) score at the recruiting station. Recruiters use CAST to help determine whether to send prospects to Military Entrance Processing Stations for further testing and to forecast the various options and benefits for which the prospects will subsequently qualify. This report summarizes analyses from nation-wide cross-validation of CAST.

This research was conducted under the Manpower and Personnel research program and contributes to the mission of the Selection and Classification Technical Area. This mission is to improve the Army's capability to select and classify its applicants' potential using state-of-the-art, fair measures. Continuing research and development of CAST is conducted under the sponsorship of the U.S. Army Recruiting Command (USAREC), as outlined in a Memorandum of Understanding regarding the ARI/USAREC Research and Development Program dated 29 August 1984. The information in this report has been briefed to the Chief of the Training Division, USAREC, in October 1985, to the Director of Recruiting Operations Directorate, USAREC, on 14 November 1986, and to the commanding officer of USAREC, MG Ono, on 7 April 1986. The results are being used to further document the acceptability of using CAST as a prescreening tool and to direct future refinement efforts.

EDGAR M. JOHNSON
Technical Director

THE COMPUTERIZED ADAPTIVE SCREENING TEST (CAST): AN EXAMINATION OF TEST VALIDITY AND TEST FAIRNESS

EXECUTIVE SUMMARY

Requirement:

To provide accurate, reliable information regarding the validity of CAST, to examine the issue of test fairness with respect to racial and sexual subgroups of examinees, and to investigate ways in which CAST could be modified to optimize its utility to recruiters.

Procedure:

A modified version of the CAST software was used in 60 recruiting stations across the country from January through December 1985 so that prospects' CAST performance could be recorded on data diskettes for analysis. The CAST scores were matched by SSN to applicant tapes from Military Entrance Processing Stations to obtain AFQT scores and relevant demographic data. These data were examined using regression and cross-tabulation analyses.

Findings:

The findings in this report are based on data gathered during the first 6 months of 1985. These analyses indicate that CAST is quite good at predicting AFQT scores for the entire sample and for examinees grouped by sex and race (black or white). When corrected for restriction in range, the cross-validated validity estimate based on the whole sample is .86. The Gulliksen-Wilks approach was used to test for subgroup differences in prediction. Subgroup differences in CAST prediction do appear to exist; however, these differences are not large and they parallel those found with most other cognitive ability tests. Analyses of CAST's accuracy at predicting prospects' classification into important AFQT categories indicate that the current version does a good job at category prediction. This report describes difficulties inherent in examining subgroup differences in category predictions.

Utilization of Findings:

The U.S. Army Recruiting Command has used these findings to support its continued use of CAST as an informal screening device and to reaffirm its decision to continue support for research and development of this test.

THE COMPUTERIZED ADAPTIVE SCREENING TEST (CAST):
AN EXAMINATION OF TEST VALIDITY AND TEST FAIRNESS

CONTENTS

LIST OF TABLES

LIST OF FIGURES

THE COMPUTERIZED ADAPTIVE SCREENING TEST (CAST): AN EXAMINATION OF TEST VALIDITY AND TEST FAIRNESS

INTRODUCTION

The Computerized Adaptive Screening Test (CAST) was developed by the Navy Personnel Research and Development Center (NPRDC) with funding from the Army Research Institute (ARI) to provide a prediction of prospective recruits' (i.e., prospects') Armed Forces Qualification Test (AFQT) scores at recruiting stations (Sands & Gade, 1983). The purpose of this report is to describe a large scale CAST data-collection effort. Analyses reported herein examine the validity of CAST and seek to address the issue of test fairness with respect to racial and sexual subgroups of examinees.

Background

Individuals interested in joining any of the armed services are required to take the Armed Services Vocational Aptitude Battery (ASVAB). ASVAB scores are used to determine eligibility for enlistment and to assist in determining initial training assignments. The ASVAB is administered under secure testing conditions either by the Department of Defense Student Testing Program or at a Military Entrance Processing Station (MEPS) or Mobile Examining Team site (METS). Most testing is conducted at MEP/MET locations. Sending individuals to these sites represents a significant financial investment for the armed services. In addition to the costs of the testing itself; travel, lodging, and boarding expenses are typically incurred. Both the recruiter and the prospect also invest a significant amount of time in this process. The recruiter must make arrangements to ensure that the prospect gets to the testing site. For the prospect, the three and one-half hours required to take the test battery must be added to the time spent getting to and from the testing location.

AFQT scores are currently computed by adding together four ASVAB subtest scores. Specifically, word knowledge (WK), arithmetic reasoning (AR), paragraph comprehension (PC) subtest scores, and one-half of the numerical operations (NO) subtest score combine to produce AFQT. An individual's AFQT score is intended to reflect his or her "trainability." Thus, AFQT scores are used to assess the eligibility of applicants for enlistment and special benefits. In the Army, applicants who score at or above the 50th percentile (AFQT categories 1, 2, and 3A) are eligible for special options and benefits such as the 2-year Enlistment Option and the Army College Fund. Applicants who score between the 31st and 49th percentiles on AFQT (AFQT category 3B) qualify for enlistment but are not eligible for special options. Those individuals who score between the 16th and 30th percentiles (AFQT categories 4A and 4B) are generally regarded as being low priority recruits.

Thus there are two major reasons why information that predicts prospects' AFQT performance is important to Army recruiters. A test that provides this information can be used simultaneously as an informal screening device and a sales tool. If the test indicates that a prospect has very little chance of subsequently qualifying for enlistment, the recruiter may choose to discourage him or her from further interest in the Army. Besides saving the expense of ASVAB testing, this allows recruiters to spend a greater amount of time selling the Army to more promising prospects. One of the major functions of a recruiter is to convince qualified prospects that the Army is a desirable job alternative. The special options and benefits offered by the Army are powerful incentives, but they only work if an individual subsequently qualifies for them. In other words, their utility depends upon the recruiter using them with the right people. Clearly, a test that predicts subsequent AFQT performance gives recruiters information they need to most effectively perform their jobs.

The Enlistment Screening Test (EST) is currently available to all of the armed services for use at recruiting stations as a predictor of AFQT performance. Although EST provides fairly accurate predictions of AFQT scores, it has several drawbacks that it shares with most other paper-and-pencil tests. The major drawbacks concern administration time, clerical burden, and scoring errors (cf. Baker, Rafacz, & Sands, 1984). Recruiters must allow prospects 45 minutes to complete EST and then they must hand score the test. This latter step takes additional time and is subject to error. Because there are currently only two alternative EST forms, it is possible that prospective applicants might learn the items and eventually pass the test on repeated testing in different recruiting stations. Excessive test time, clerical burden, and test security are problems that can be alleviated or eliminated because of recent advances in computer technology and psychometric theory.

An advance in psychometric theory, called Item Response Theory (IRT), has made it possible to adapt or "tailor" a test to the individual examinee (Lord, 1980). Unlike ability tests based on classical test theory, ability tests based on IRT can provide comparable estimates of individuals' ability levels even when different individuals receive different sets of test items. Adaptive testing makes it possible to construct tests that are able to discriminate equally well across all ability levels. In addition to improving the discriminability of a test, computerized adaptive tests are more efficient to use than traditional paper-and-pencil tests because they reduce testing time without sacrificing validity. Computerized adaptive tests also eliminate the need for manual scoring and recording which can result in clerical errors, and they can provide immediate feedback on test results. Computerized adaptive tests reduce test compromise by eliminating test booklets which can be stolen, and by administering different items to different individuals making it more difficult for individuals to "cheat." For all of these reasons, a computerized adaptive test that can accurately predict a prospect's AFQT score is a highly desirable recruiting tool; thus, the Computerized Adaptive Screening Test (CAST) was developed.

In 1983, the U.S. Army Recruiting Command (USAREC) procured a microcomputer system that would lend invaluable assistance in the achievement of the recruiting mission. The system, known as the Joint Optical Information Network (JOIN), consists of a microprocessor, printer, television monitor, video disk player, and modem. JOIN was designed to serve a number of uses in recruiting stations and MEPS. It can be used to introduce the Army to those individuals who may be interested in enlisting by providing information via videodisks. For example, there are video segments that show recruits in basic training, soldiers in various Army jobs, and the enlistment options and benefits that the Army provides. In the near future, JOIN will also be used to implement interactive recruiter training (for self-paced, on-site professional development) and to forward data regarding Army applicants to a central database. For the time being, however, one of the primary advantages of having JOIN is that it has allowed the Army to be among the first employers to take full operational advantage of the benefits of computerized adaptive testing.

Development of CAST

The item pools for CAST were constructed by researchers at the University of Minnesota (cf. Moreno, Wetzel, McBride, & Weiss, 1984) in early experimental work on a computerized adaptive version of ASVAB (called CAT ASVAB). Moreno et al. provided a de facto pilot test of CAST in their research which examined the relationship between corresponding ASVAB and CAT ASVAB subtests. These researchers administered the WK, AR, and PC subtests to 270 male Marine recruits at the Marine Corps Recruit Depot in San Diego, California. The data from this pilot test yielded a correlation of .87 between the three optimally-weighted CAT ASVAB subtests and ASVAB AFQT. Because the statistical analyses indicated that the PC subtest did not contribute a significant amount of predictive power beyond that provided by the WK and AR subtests, and because the PC subtest items required an inordinate amount of time to administer, this subtest was dropped from CAST.

Presently, there are 78 items in CAST's WK item pool and 225 items in CAST's AR item pool. All items are multiple choice with a maximum of five response alternatives. CAST uses a three-parameter logistic ogive item response model (Birnbaum, 1968); thus each item has three parameters (discrimination, difficulty, and guessing) associated with it. Test items for CAST item pools were chosen so that the discrimination parameter values would be greater than or equal to .78; the difficulty parameter values would range between +2 and -2; and the guessing parameter values would be less than or equal to .26. CAST uses the Bayesian sequential scoring procedure discussed by Jensema (1977) to score and select subsequent items for administration. The test ends when the examinee has responded to 10 WK and 5 AR items.

Prior Validation Efforts

There are three validation efforts associated with CAST. The initial validation project was conducted at the Los Angeles MEPS with a sample of 312 U.S. Army applicants (Sands & Gade, 1983). Each applicant received 20

WK items and 15 AR items on an APPLE-II microcomputer. The data were analyzed to determine the optimal combination of subtest lengths so that the predictive accuracy of CAST would be at least as high as that estimated of EST (r=.83; Mathews & Ree, 1982) with the shortest administration time possible. Multiple correlation coefficients were computed for each of the 300 combinations of subtest lengths. Examination of the results led to the recommendation that the operational version of CAST be terminated following the administration of 10 WK and 5 AR items. The multiple correlation between this optimally-weighted subtest score combination and actual AFQT score was .85.

Army recruiting stations in the midwestern region of the U.S. provided CAST cross-validation data during January and February, 1984 (Pliske, Gade, & Johnson, 1984). CAST was introduced by geographical region, and the midwestern region was the only fully operational region at the time of data collection. Recruiters in these stations recorded prospects' CAST scores and social security numbers (SSNs) on log sheets. USAREC collected these data and forwarded them to ARI for analysis. The CAST scores recorded by the recruiters were matched by SSNs to applicant data tapes supplied by the Military Entrance Processing Command (MEPCOM) to obtain AFQT scores and relevant demographic data. Matching records were located for 1,962 individuals. The bivariate correlation coefficient between CAST and AFQT scores computed from these data was .80. This value reflects a reasonable amount of shrinkage from the original validity estimate of .85.

Although the validity estimates yielded by these two projects suggest that CAST is an effective predictor of AFQT, an additional data collection effort was required. Two goals of this data collection effort would be (1) To provide highly accurate, stable information regarding the validity of CAST, and (2) To examine the issue of test fairness with respect to racial and sexual subgroups of examinees. Thus, a large scale cross-validation effort using a sample representative of all Army prospects was called for.

METHOD

Data Collection Procedure

Currently, the JOIN system is programmed to record background information and CAST percentile scores onto Prospect Data diskettes. Because recruiters are required to maintain this information for only as long as they need it, and because more detailed information was required, a modified version of the CAST software was designed for use in this latest validation project. The program was changed so that examinees' test responses would be recorded on special data collection floppy diskettes that could be forwarded to ARI for analysis. Information recorded on the data diskettes included item identification number, examinee's answer, the time it took for the examinee to read and answer the item, and the examinee's SSN. The software was also changed so that the prospects would respond to

15 WK and 10 AR items. However, the predicted AFQT score reported at the end of the test was based on the operationally-used stopping rule of 10 WK and 5 AR items.

The modified CAST software was distributed for use in 60 recruiting stations located across the country. These stations were selected to be representative of the population of approximately 2,000 Army recruiting stations in terms of geographic location and population density. A full year of data collection was required to ensure that the sample of prospects would not be biased by seasonal fluctuations in prospect characteristics. The analyses discussed in this paper are based on data collected during the first six months of this project.

Army recruiters use EST rather than CAST when they do not have access to their JOIN systems. Because only one EST validity estimate has been reported, this seemed to be an ideal opportunity to collect cross-validation data. Consequently, the 60 participating recruiting stations were also given log sheets to record the scores of prospects to whom they administered EST. In addition to recording the raw EST scores, the recruiters were asked to record the prospects' SSNs. The EST log sheets were forwarded to ARI along with the CAST data diskettes at the end of each month. The CAST and EST scores recorded at the recruiting stations were matched to ASVAB and demographic data available on computer tapes supplied by MEPCOM. The focus of the present report is on the cross-validation of CAST. Preliminary analyses of the EST data are provided in Knapp and Pliske (1986).

Sample Characteristics

Table 1 summarizes the major demographic characteristics of the CAST sample. It is difficult to determine the extent to which the sample accurately represents the population of Army prospects because no data are available to describe that population accurately. It is likely, however, that the sample exhibits differences from the Army prospect population because many prospects fail to go to MEPS for ASVAB testing and this sample is based only on those prospects for whom we located a matching MEPCOM record. On the basis of the information provided to them by recruiters, some prospects decide that they are not interested in joining the Army so they do not go to MEPS. Further, recruiters choose not to encourage some prospects to go to MEPS because their prequalification information suggests that the prospects are unsuitable for enlistment in the Army. Thus certain kinds of prospects are being systematically excluded from the sample. One result of this situation is that there is a restriction in the range of CAST scores. Given the absence of more appropriate criteria, the adequacy of other characteristics of this sample can be evaluated in terms of the sample selection procedure and common sense expectations. The sampled recruiting stations were selected to be representative of all recruiting stations in terms of geographical location and population density. Because blacks represent a small percentage of the population of American citizens, the

Table 1

National CAST Cross-validation (January-June 1985)
Sample Description

Sample Size	2,214
Sex	81% Male
	19% Female
Race	59% White
	37% Black
	4% Other
Age	Mean=20; SD=3.47
	Median=19
	Mode=18
Component	85% Regular Army
	15% Army Reserve
Education	4% Some College/Vo Tech
(Based on 65% Cases)	77% HS Diploma or GED
	19% Non-HS Graduates[a]
AFQT Category	25% 1 and 2
(From ASVAB)	16% 3A
	28% 3B
	31% 4A-5

[a] Includes high school seniors

sample selection procedure was also designed to insure that a relatively large number of black prospects would be included. A sufficient number of black prospects is needed to permit legitimate comparison to white prospects, a major goal of this project. Other characteristics (e.g., average age and percentage of males) of the sample correspond quite well with a priori expectations.

Analytical Procedures

To address the issue of test validity, several analytical approaches were taken. Besides computing the simple b.variate and corrected bivariate correlations between CAST and AFQT[1] scores, the regression lines were computed and displayed to provide a better understanding of the nature of

CAST's linear AFQT predictions. The validity of CAST was also examined with respect to its ability to predict whether examinees will fall above or below important AFQT cutpoints (i.e., category predictions).

The second goal of this research project has been to examine the issue of test fairness. The most commonly used statistical analysis that addresses this concern is based on the linear regresssion model, and is known as the Gulliksen and Wilks test (Gulliksen & Wilks, 1950). This test for subgroup differences in prediction is a three-step process that begins with a comparison of standard errors of estimate. This comparison is performed first because regression lines that exhibit different standard errors of estimate can not be directly compared. Specifically, this chi-square test is sensitive to differences in the amount of variance in the predictor data. If subgroups do not exhibit significant differences in the size of their standard errors, then the slope and the intercept of their regression equations can be tested for significant differences in size across subgroups. Although the relative size of the intercepts and slopes are not directly comparable across subgroups, the size of the intercept is related to average performance on the criterion (AFQT) and the size of the slope indicates the change in the criterion associated with one unit of change in the predictor. The statistical tests for differences in slopes and intercepts can be performed simultaneously and results are reported using the $\underline{F}$ statistic. Generally speaking, if the regression equations for each subgroup of examinees are equal then the use of a common regression equation for actual prediction is warranted. Presumably, the common regression equation would then result in optimal prediction for all examinees and no differences among subgroups in the degree of overprediction or underprediction of subsequent performance would occur.

Test fairness is also examined by performing a stepwise regression analysis that shows the extent to which variables such as sex, age, race, and the version of ASVAB that was administered to the examinee add to CAST's power to predict AFQT. In addition, the accuracy of CAST's category predictions are broken down by race and sex. No inferential statistics were computed, however, and no attempt was made to compare statistically the relative accuracy of subgroup predictions. We are not aware of any way to equate the different groups with respect to variance in CAST performance or their base rates, so accuracy rates are not comparable across the groups.

The results that follow are presented in two major sections: Linear predictions and category predictions. Because test fairness is so closely associated with test validity, the relevant analyses and comparisons will be subsumed under the two major validity sections. It is important to note that only data from white and black subgroups were used in analyses dealing with race because members of other ethnic groups were not adequately represented in the sample.

---

[1] In the analyses reported herein, AFQT scores are based on corrected 1980 Youth Norms.

RESULTS

Linear Predictions

The CAST validity estimates from the present investigation are shown in Table 2. Because there is some degree of range restriction in the CAST scores, it is appropriate to correct the correlations for this statistical artifact. The corrections are based on the standard deviation of all the CAST scores (including those scores that could not be matched to ASVAB records) that were forwarded to ARI January through June, 1985 (N=6,470; Mean=40.36; SD=21.85). The corrected correlation between CAST and AFQT for the entire sample is .86. There are two other important points to note about the information in this table. First, correcting the correlations for restriction in range greatly decreases the difference between the white and black subgroups. Second, the fact that the corrected correlation for the white subgroup is still a little larger than that for the black subgroup is probably due to the fact that the common prediction equation is based on a larger number of white than black examinees.

Table 2

Bivariate Correlation between CAST and AFQT Scores
by Race and Sex

Group	$\underline{r}$	$\underline{r}_c$[a]
All	.82	.86
White	.81	.85
Black	.70	.82
Male	.82	.85
Female	.81	.86

[a] Correlations corrected for restriction in range of CAST scores.

Table 3 shows the regression equations that describe the prediction of AFQT scores from CAST scores for the entire sample and for examinees grouped by race and sex. In addition to reporting the squared correlation coefficients, the standard errors of estimate (SEest's) are shown. The standard error of estimate, like the correlation coefficient, reflects the strength of the relationship between the predictor and criterion. Unlike the raw correlation coefficient, however, the size of the standard error of estimate is partially determined by the amount of variance in the predictor scores.

Table 3

Regression of AFQT Scores onto CAST Scores

Subgroup	N	Intercept	Slope	r^2	SEest
All	2,214	-1.19	1.04	.67	13.74
White	1,309	.04	1.05	.65	13.90
Black	816	3.68	.83	.49	12.79
Male	1,795	-2.29	1.04	.68	13.70
Female	419	1.97	1.05	.66	13.33

Recall that the Gulliksen and Wilks test for differences in subgroup prediction begins with a comparison of the respective standard errors of estimate. With regard to the racial subgroups, the standard error of estimate associated with the regression equation for black examinees is lower than that associated with the white subgroup (chi-square=7.94, df=1).[2] This means that, if separate prediction equations were used, the prediction of black AFQT performance would be slightly more accurate than the prediction of white AFQT performance. Because the comparison of standard errors of estimate indicates that there are statistically significant racial differences in prediction, there is no justification for proceeding with a comparison of the slopes and intercepts of the two equations.

There is not, however, a significant difference between the standard errors of estimate associated with the male and female subgroups (chi-square=.6, df=1). Therefore, a simultaneous test for intercept and slope differences was performed. This test indicated that there is a statistically significant difference between the prediction equations of the two subgroups ($\underline{F}$=20, df=2, 2210).

A visual representation of the subgroup regression lines will aid in the evaluation of the subgroup differences detected here. Figure 1 shows the regression lines for two racial subgroups and the common regression line that is based on the entire sample (including other racial groups). The regression line for white examinees closely parallels the common regression line, but it lies a little above the common line. This indicates that white AFQT performance is subject to underprediction when a common regression line is used for prediction. Black examinees are underpredicted at the low end of the AFQT continuum (below the 20th percentile) but they are overpredicted across the rest of the continuum. Underprediction occurs when performance on the predictor test suggests that examinees will perform worse on the criterion than they actually do. Overprediction occurs when performance on the predictor indicates that examinees will perform better on the criterion than they actually do.

----

[2] Throughout this report, the level of statistical significance is $\underline{p} < .01$.

DEPICTION OF CAST REGRESSION BY RACE

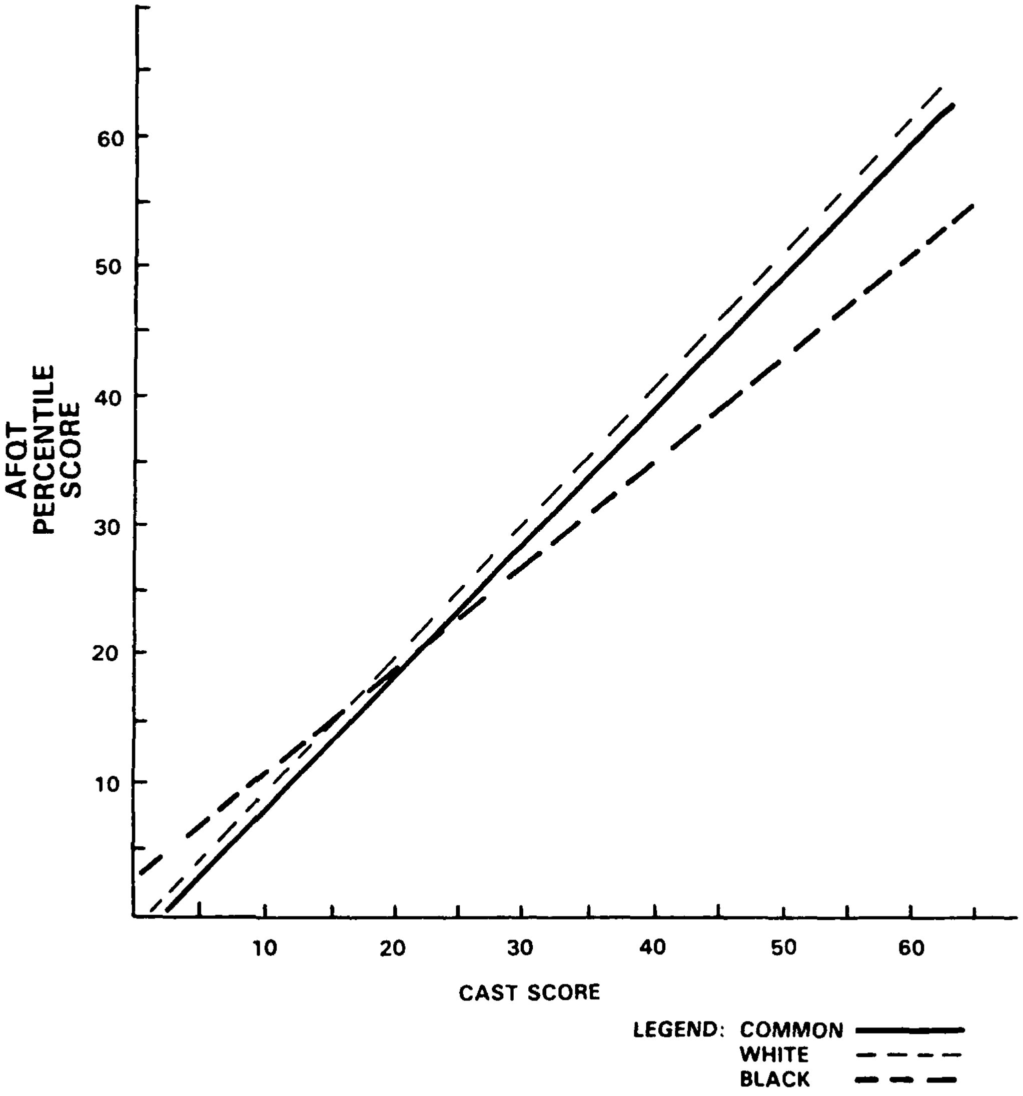

Figure 1

Figure 2 displays the common regression line and the regression lines computed for male and female examinees separately. Use of a common regression line results in the underprediction of female AFQT performance and the overprediction of male AFQT performance. The lines are roughly parallel, but they differ in elevation. Differences in elevation reflect differences in average AFQT performance.

Actual performance differences between racial and sexual subgroups are shown in Table 4. These differences reflect the conclusions made on the basis of the tests of differences between regression lines and the depictions of those regression lines in Figures 1 and 2. Theoretically, percentile scores are rectangularly distributed, however in this sample they more closely resemble a normal distribution. Therefore, Table 4 also shows the standard deviations associated with the score distributions.

Table 4

Mean Test Performance by Race and Sex

	AFQT		CAST	
Subgroup	Percentile Score	SD	Score	SD
All (n=2,214)	47	23.9	46	18.8
White (n=1,309)	55	23.6	52	18.2
Black (n=816)	34	17.9	36	15.2
Male (n=1,795)	46	24.1	47	19.1
Female (n=419)	48	22.8	44	17.6

A stepwise multiple regression analysis was conducted to examine the impact of factors other than CAST performance that may increase the predictability of AFQT scores. Table 5 summarizes the results of this analysis. The only variables that add to the predictive power of CAST are sex and race (either Black or White). The increase in explained variance due to these factors is small, corroborating the evidence that has already been presented.

DEPICTION OF CAST REGRESSION BY SEX

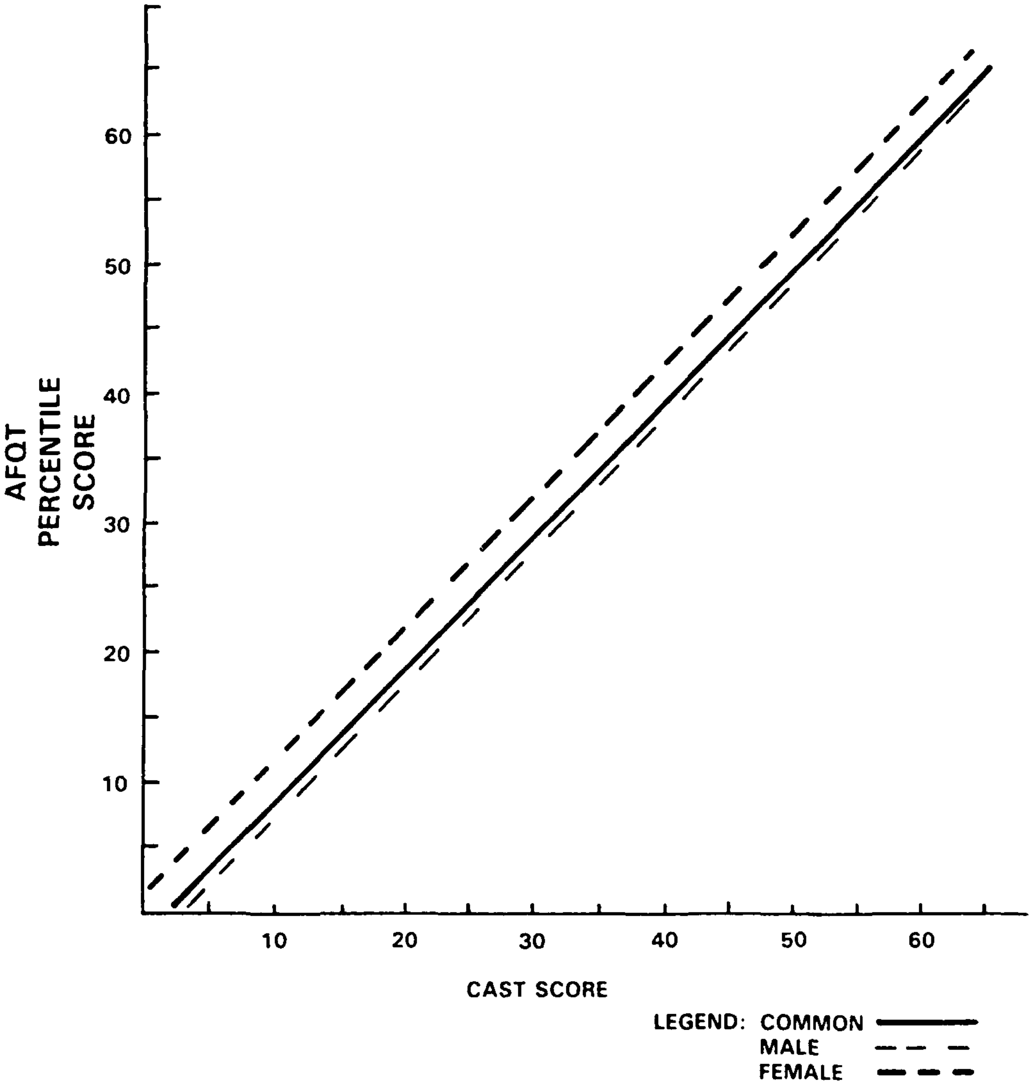

Figure 2

Table 5

Percent Variance Accounted for by Stepwise Addition of Variables to Regression Model

Predictor	R^2
CAST Score	.668
Race	.678
Sex	.686
Years of Education	.688
Age	.690
ASVAB Version	.690

To summarize, CAST has a strong linear relationship with AFQT. There are minor differences across subgroups of examinees with respect to the nature of CAST's predictions when a common regression line is used. To the extent that the predictions are in error, the AFQT performance of black examinees tends to be overpredicted whereas the AFQT performance of white examinees tends to be underpredicted. With respect to a comparison between the predictions of males and females, the performance of female performance is underpredicted relative to the overprediction of male AFQT performance. Although the differences in subgroup prediction are statistically significant, they are not large.

Category Predictions

At the present time, CAST provides feedback in the form of bar charts that represent examinee performance on the WK and AR subtests and the examinee's predicted AFQT percentile score. The great majority of recruiters, however, have never been taught the fundamentals of regression analysis, and thus, do not adequately understand the nature of point predictions. Hence, recruiters expect predicted and actual AFQT scores to be exactly the same. Further, Army recruiters are primarily interested in making category predictions rather than point predictions. Specifically, they first want to know whether a prospect will likely qualify as a desirable Army enlistee (i.e., scoring above the 30th percentile so that he or she will be classified into AFQT category 3B or above). Secondly, they want to know if the prospect will likely qualify for special options and benefits that are available if he or she scores at or above the 50th AFQT percentile (i.e., AFQT category 3A or above). Given this situation, it may be more useful to provide the recruiters with probabilistic predictions tied to subsequent category classifications. For example, the prediction interval associated with the point prediction could be portrayed alongside an AFQT percentile score continuum on which important category cutpoints are highlighted. Alternatively, the estimated probability that the examinee will subsequently be classified into category 3B or above on AFQT and the probability that the examinee will susequently be classified into AFQT category 3A or above could be reported by CAST.

Currently, most recruiters probably make AFQT category predictions by interpreting CAST scores at face value. For example, a prospect who receives a predicted AFQT score of 31 would be predicted to belong to AFQT category 3B and a prospect with a predicted AFQT score of 30 would be predicted to belong to AFQT category 4A. Assuming that this is the way in which recruiters use CAST scores, this prediction scenario can be modeled statistically.

Figure 3 shows CAST prediction results at the 3B/4A cutpoint (i.e., the 31st AFQT percentile) when the aforementioned assumption is made. Out of the entire CAST sample, a total of 81% of the examinees are correctly classified into either the 1-3B category (65%) or the 4A-5 category (16%). The performance of most of the examinees misclassified by CAST was overpredicted. That is, when CAST was wrong, it was most likely to misclassify an unqualified examinee into the "passing" category. Figure 3 also shows the CAST prediction results at the 3A/3B cutpoint (i.e., the 50th AFQT percentile). In this case, errors are slightly more likely to be underpredictions (9%) rather than underpredictions (8%). The overall hit rate (i.e., the percentage of correct predictions) is 83%.

Prediction results at the two cutpoints for the black and white subgroups are shown in Figure 4. At the 31st percentile, the white examinees are correctly classified more often than the black examinees (87% hit rate versus 70% hit rate). At the 50th percentile, it appears that the black exmainees are correctly classified more often than the white examinees (87% hit rate versus 80% hit rate). However, unlike the standard error of estimates and the corrected correlation coefficients reported in earlier analyses, the hit rates shown here are not directly comparable because they ignore subgroup differences in predictor variance. Specifically, because predictor variance for the black subgroup is smaller than the variance for the white subgroup, the predictive accuracy (i.e., hit rate) is artificially constrained. Therefore, it is likely that subgroup differences in hit rate would be alleviated or eliminated if the subgroups exhibited equal variances in CAST performance. The information in Figure 4 indicates that errors of overprediction and underprediction are roughly similar for both subgroups. This is not immediately obvious because the values presented in the tables must be adjusted for subgoup differences in hit rates. This can be done by dividing the total percentage of inaccurate predictions (i.e., "misses") by the percentage of underpredictions (or conversely, overpredictions) for each subgroup. This exercise shows that very small racial differences with respect to the magnitude of the prediction errors exist. At the 3B/4A cutpoint the performance of 23% are underpredicted for both subgroups. At the 3A/3B cutpoint the performance of white examinees misclassified by CAST is subject to slightly more underprediction than their black counterparts (55% versus 50%).

Male and female examinees are compared in Figure 5. The hit rates at both the 3A/3B and 3B/4A cutpoints are quite similar for male and female examinees. At the 3B/4A cutpoint, males have an overall hit rate of 81% as compared to females who have an overall hit rate of 78%. At the 3A/3B sutpoint, males have an overall hit rate of 82% as compared to females who

PATTERN OF CAST PREDICTIONS AT TWO AFQT CATEGORY CUTPOINTS

		CAST SCORE	
		BELOW 31	31 OR ABOVE
AFQT PERCENTILE SCORE	31 OR ABOVE	4% Underprediction	65% Hit
	BELOW 31	16% Hit	15% Overprediction

Figure 3a

		CAST SCORE	
		BELOW 50	50 OR ABOVE
AFQT PERCENTILE SCORE	50 OR ABOVE	9% Underprediction	32% Hit
	BELOW 50	51% Hit	8% Overprediction

Figure 3b

*Note that the percentages in each table total 100;
Percentiles based on corrected 1980 norms.

PATTERN OF CAST PREDICTIONS AT TWO AFQT CATEGORY CUTPOINTS

BY RACE

AFQT PERCENTILE SCORE	BELOW 31	31 OR ABOVE
31 OR ABOVE	3%	79%
BELOW 31	8%	10%

CAST SCORE
WHITE

AFQT PERCENTILE SCORE	BELOW 31	31 OR ABOVE
31 OR ABOVE	7%	41%
BELOW 31	29%	23%

CAST SCORE
BLACK

Figure 4a

AFQT PERCENTILE SCORE	BELOW 50	50 OR ABOVE
50 OR ABOVE	11%	44%
BELOW 50	36%	9%

CAST SCORE
WHITE

AFQT PERCENTILE SCORE	BELOW 50	50 OR ABOVE
50 OR ABOVE	6.5%	12%
BELOW 50	75%	6.5%

CAST
BLAC

Figure 4b

*Note that the percentages in each table total 100; Percentiles based on corrected 1980 Youth Norms.

PATTERN OF CAST PREDICTIONS AT TWO AFQT CATEGORY CUTPOINTS*

BY SEX

AFQT PERCENTILE SCORE		BELOW 31	31 OR ABOVE
	31 OR ABOVE	4%	65%
	BELOW 31	16%	15%

CAST SCORE
MALE

AFQT PERCENTILE SCORE		BELOW 31	31 OR ABOVE
	31 OR ABOVE	8%	64%
	BELOW 31	14%	14%

CAST SCORE
FEMALE

Figure 5a

AFQT PERCENTILE SCORE		BELOW 5Ø	5Ø OR ABOVE
	5Ø OR ABOVE	9%	32%
	BELOW 5Ø	5Ø%	9%

CAST SCORE
MALE

AFQT PERCENTILE SCORE		BELOW 5Ø	5Ø OR ABOVE
	5Ø OR ABOVE	13%	29%
	BELOW 5Ø	54%	4%

CAST SCORE
FEMALE

Figure 5b

*Note that the percentages in each table total 1ØØ; Percentiles based on corrected 198Ø Youth Norms.

have a hit rate of 83%. When the slight difference in hit rates is taken into account, it can be seen that females are more likely than males to have their performance on AFQT underpredicted. Of those female examinees misclassified by CAST at the 31st percentile, the performance of 36% was underpredicted compared to 21% for males. Of those female examinees misclassified by CAST at the 50th percentile, the AFQT performance of 76% was underpredicted. The corresponding figure for male examinees is 50%.

The preceding analyses on CAST's category predictions corroborate the analyses presented on CAST's linear predictions. They show that CAST has a general tendency to overpredict the AFQT classification of prospects at the 3B/4A cutpoint. Although differences in predictor variance warrant caution in interpreting such comparisons, race and sex differences in predictive accuracy do not appear to be large. With respect to prediction errors, these analyses also support and clarify the linear analyses that indicated that these errors vary across the ability continuum. For the most part, black examinees are somewhat more likely than white examinees to have their subsequent performance on AFQT overpredicted. The AFQT performance of female examinees, relative to male examinees, tends to be underpredicted.

DISCUSSION

In January 1985, 60 Army recruiting stations were asked to begin forwarding CAST and EST data to ARI. Although this data collection effort continued through December 1985, only the data collected during the first six months of 1985 have been analyzed. Based on these analyses, it can be concluded that the current operational version of CAST is reasonably accurate at predicting AFQT scores ($\underline{r}_c$ =.86).

The topic of test fairness has received a great deal of attention in the research literature of the past ten years. Although there is still a substantial amount of disagreement as to what constitutes an unfair test, the relevant issues have emerged more clearly in recent years. Messick (1975) distinguished two questions that should be asked about any given test. The first question asks whether the test is valid. That is, does the test adequately measure what it purports to measure? This is a technical question that is best answered by psychometric experts. The second question asks whether the intended use of the test would serve public interests. For example, a selection test that validly screens out disproportionate numbers of racial minorities may be judged detrimental to society's goal of minimizing the effects of racially discriminatory labor practices. This latter issue is clearly a question of policy.

The failure to distinguish the nature of the two questions described above led to a general state of confusion in the testing literature during the 1970's. Many testing experts simply chose to equate test fairness with test validity. This is not necessarily the wrong approach. Indeed, it is the approach that was taken here. What is important to recognize, however, is that the use of this approach implies certain policy decisions that are often not made explicit. The analyses reported herein, and the conclusions drawn on the basis of those analyses, entail the following assumptions of policy. Stated simply, correctly predicting that a person will "pass" ASVAB AFQT has a positive utility equal to the utility associated with correctly predicting that a person will fail ASVAB AFQT. (Note that the notion of passing or failing AFQT is introduced here to simplify this discussion.) Overpredicting subsequent AFQT performance and underpredicting subsequent AFQT performance are considered to be outcomes of no utility. Finally, the utilities associated with each of the four possible outcomes (i.e., correct acceptance, correct rejection, false positive, and false negative) do not vary as a function of race or sex. For example, underpredicting the performance of white prospects is judged to be as undesirable as underpredicting the performance of black prospects. These policy assumptions correspond to the definition of test fairness that is most often used in the testing literature. Cleary (1968) states that "...a test is biased if the criterion score predicted from the common regression line is consistently too high or too low for members of [a] subgroup." (p. 115). Cleary's definition of unfair test bias can be described another way. A test is biased if the percentage of examinees overpredicted or underpredicted significantly differs by racial or sexual subgroups.

Race and sex subgroup differences in AFQT predictions based on CAST exist, but the magnitude of these differences is not unreasonably large. The pattern of differences parallels those found in validity studies of ASVAB (e.g., Dunbar & Novick, 1985; Hanser & Grafton, 1983) and of college entrance examinations like the Scholastic Aptitude Test (e.g., Kallingal, 1971; Temp, 1971). The data show that regression lines (particularly the intercepts) are likely to differ across subgroups, and the use of a common regression line tends to favor black and, to a far smaller extent, male applicants. That is, the common regression line tends to overpredict the performance (or qualifications) of those subgroups.

In the present report, it was shown that correcting for restriction in the range of CAST scores alleviates apparent subgroup differences in prediction. These results (which were based on the analysis of continuous data) allowed for the assumption that the racial differences in range restriction that created apparent racial differences in AFQT score prediction were also responsible for creating apparent racial differences in AFQT category predictions. If only category data were available, however, the assumption as to the extent to which range restriction differences account for differences in prediction would be more tenuous. The authors are not aware of any acceptable statistical approach to this problem.

The present CAST validation effort, and those preceding it, have used a criterion-related validation paradigm. This validation approach is vital to the evaluation of a test such as CAST. Showing that a test predicts what it is supposed to predict, however, is not sufficient for showing that the test measures what it is supposed to measure (construct-related validity). CAST attempts to measure two underlying abilities: Word knowledge and arithmetic reasoning. The item calibration procedures used in Item Response Theory methodology are intended to insure that individuals with the same level of the ability have the same probability of answering a given test item correctly. When the test items that currently compose CAST's item pools were calibrated, however, the calibration procedure was performed on all examinees simultaneously. Because items were not also calibrated on examinees grouped by sex and race, there may be items that exhibit construct validity with respect to one subgroup but not another. For example, an item would be racially biased if blacks at a particular ability level are less likely than whites at the same ability level to get the item correct.

There are plans to refine CAST in the near future. New test items will be developed to ensure the continued intregrity of the test. The new items and currently existing items will be calibrated on both white and black subgroups of examinees. The refinement effort will also entail a reevaluation of CAST's item selection rules and scoring algorithm. Given the desire to optimize prediction at certain AFQT category cutpoints, the refinement effort will be directed toward this goal.

REFERENCES

Baker, H. G., Rafacz, B. A., & Sands, W. A. (1984). *Computerized Adaptive Screening Test (CAST): Development for use in military recruiting stations* (NPRDC Report No. 84-17). San Diego, CA: Navy Personnel Research and Development Center.

Birnbaum, A. (1968). Some latent trait models and their use in interfering an examinee's ability. In F. M. Lord and M. R. Novick (Eds) *Statistical theories of mental test scores*. Reading, Mass: Addison-Wesley.

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement*, *5*, 115-124.

Dunbar, S. B. & Novick, M. R. (1985). *On predicting success in training for males and families: Marine Corps clerical specialties and ASVAB forms 6 and 7* (ONR Report No. 85-2). Washington, DC: Office of Naval Research.

Gulliksen, H. & Wilks, S. S. (1950). Regression tests for several samples. *Psychometrika*, *15*, 91-114.

Hanser, L. M. & Grafton, F. C. (1982). *Predicting job proficiency in the Army: Race, sex, and education* (Selection and Classification WP No. 82-1). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Jensema, C. G. (1977). Bayesian tailored testing and the influence of item bank characteristics. *Applied Psychological Measurement*, *1*, 111-120.

Kallingal, A. (1971). The prediction of grades for Black and White students at Michigan State University. *Journal of Educational Measurement*, *8*, 263-266.

Knapp, D. J. & Pliske, R. M. (1986). *Preliminary report on a national cross-validation of the Computerized Adaptive Screening Test (CAST).* ARI Research Report No. 1430, Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. AD A175 767

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Mathews, J. J. & Ree, M. J. (1982). *Enlistment Screening Test Forms 81A and 81B; Development and Calibration* (AFHRL Report No. 81-54). Brooks Air Force Base, Texas: Air Force Human Resources Laboratory.

Messick, S. (1975). The standard problems: Meaning and values in measurement and evaluation. *American Psychologist*, *30*, 955-966.

Moreno, K. E., Wetzel, C. D., McBride, J. R., & Weiss, D. J. (1984). Relationship between corresponding Armed Services Vocational Aptitude Battery (ASVAB) and computerized adaptive test (CAT) subtests. Applied Psychological Measurement, 8, 155-163.

Pliske, R. M., Gade, P. A., & Johnson, R. M. 1984). Cross-Validation of the Computerized Adaptive Screening Test (CAST) (ARI Research Report No. 1372). Alexandria, VA; U.S. Army Research Institute for the Behavioral and Social Sciences. AD A163 148

Sands, W. A. & Gade, P. A. (1983). An application of computerized adaptive testing in U.S. Army Recruiting. Journal of Computer-Based Instruction, 10, 87-89.

Temp, G. (1971). Validity of the SAT for Blacks and Whites in thirteen integrated institutes. Journal of Educational Measurement, 8, 245-252.